

Détection de fraudes et loi de probabilité de Newcomb-Benford

FERNANDEZ Christelle

PONCHEELE Clément

EL KAÏM Laura

Encadré par M.DUCHARME

24 mai 2021

Table des matières

Introduction.	3
Naissance de la loi de Newcomb-Benford.	4
Expérimentation sur différents jeux de données.	6
La suite de Fibonacci.	6
Nombres extraits d'un magazine et d'un journal.	7
Population des villes de France.	8
Passage journalier de vélos dans l'allée Beracasa à Montpellier.	9
Nombres générés par les humains	9
Tests	11
Généralités sur les tests	11
Hypothèse nulle et hypothèse alternative	11
Les risques d'erreurs	12
Test du Khi-Deux	13
Test de Freedman-Watson	14
Tests lisses pour la Loi de Newcomb-Benford	15
Application des tests	16
CAS COVID	20
Bibliographie	20

RESUME DU PROJET EN QLQ LIGNES

Mots-clés : Loi de Benford-Newcomb, 1^{er} chiffre significatif, Test d'hypothèses , Risque d'erreur.

REMERCIEMENTS

Introduction.

La fraude est une pratique répandue dans de nombreux domaines comme par exemple la finance, le secteur social ou médical. Il peut être tentant pour un être humain ou une société de tricher si cela peut impliquer pour lui une position plus confortable dans la société, telle qu'une réduction de charges, ou même un avantage sur un de ses concurrents. Il semblerait donc logique que des personnes cherchent à déceler ces fraudes.

Les données transmises par un individu ou un organisme peuvent faire l'objet de modifications, c'est de ce type de fraudes auquel nous nous intéresserons ici, et plus particulièrement la modification du premier chiffre significatif (le premier chiffre d'un nombre qui n'est pas un zéro) de nombres pris dans un certain ensemble de données.

De telles modifications entraînent un changement de la répartition des chiffres présents naturellement¹. Si nous connaissons la répartition des chiffres présentés dans un ensemble de données arbitraires, il est donc techniquement possible de savoir si un nombre a été modifié ou non.

Il nous vient donc les questions suivantes : *Qu'elle est cette répartition ? Est-il possible de la connaître et si oui, dans quels cas ?*

De manière intuitive nous pourrions penser que les nombres sont répartis de manière uniforme. Qu'en est-il vraiment ?

La première partie de notre projet consistera à **répondre à ces questions**, nous nous appuierons sur les travaux de Simon Newcomb et Frank Benford, qui ont théorisé la **loi de Newcomb-Benford**, plus communément appelée loi de Benford. Cette loi nous dit que, dans une liste de données dites naturelles, la probabilité d'avoir le chiffre i comme premier chiffre significatif est de $\log_{10}(1 + \frac{1}{i})$.

Par exemple, le chiffre 1 en tant que premier chiffre significatif serait présent à hauteur de 30% alors que le 9 à seulement 4,6%.

Dans la suite **nous mettrons en œuvre une série d'expérimentation** pour constater ou non la véracité de cette loi, pour ce faire dans un premier temps nous récolterons des nombres pris dans des milieux sensés satisfaire la loi de Newcomb-Benford et observerons la répartition du premier chiffre significatif. Puis nous répliquerons une version simplifiée de l'expérience de Hill (1988), qui consiste à observer la répartition du premier chiffre significatif d'une liste de nombre donnée au hasard par des êtres humains, en l'occurrence ses élèves.

Cette expérience est à la base des méthodes de détection de fraudes par la loi de Newcomb-Benford. Si un fraudeur modifie un jeu de données, ce jeu est donc influencé par la pensée humaine, il ne suit donc plus la loi de Newcomb-Benford. Pour détecter la fraude il suffit donc de comparer les premiers chiffres significatifs. Cependant ces comparaisons doivent se faire de manière rigoureuses et scientifiques. Pour cela il existe des test statistiques, dont le plus connu, le test du χ^2 , ou bien celui de Ducharme et collab. (2020).

Il nous vient donc les questions suivantes : *Ces tests sont-ils fiables ? Existe-t-il un test significativement meilleur que les autres ? Vont-ils dans le même sens ? Et sinon que faire ?*

La réponse à ces question constituera donc la deuxième partie de ce projet, pour ce faire nous mettrons en œuvre différents tests sur des jeux de données comme la fiscalité italienne.

1. Les données dites naturelles sont celles qui n'ont pas été influencé par la pensée de l'homme.

Naissance de la loi de Newcomb-Benford.

Il serait tentant de penser que les nombres sont répartis de manière uniforme, cela viendrait du biais d'équiprobabilité². Ce dernier consiste à “penser qu'en l'absence d'information, tous les cas ont la même probabilité de se produire et que le hasard implique nécessairement l'uniformité”.

Néanmoins cette hypothèse sera contredite une première fois par l'astronome, mathématicien, économiste et statisticien canadien Simon Newcomb. Ce dernier fournira en 1881 une première approche au principe statistique, qui se fera injustement appeler *Loi de Benford*. Celui-ci remarquera que les premières pages des tables logarithmiques sont plus utilisées que les pages suivantes (cf Figure 1). Il publiera sa découverte dans un article de l’*“American Journal of Mathematics”*.

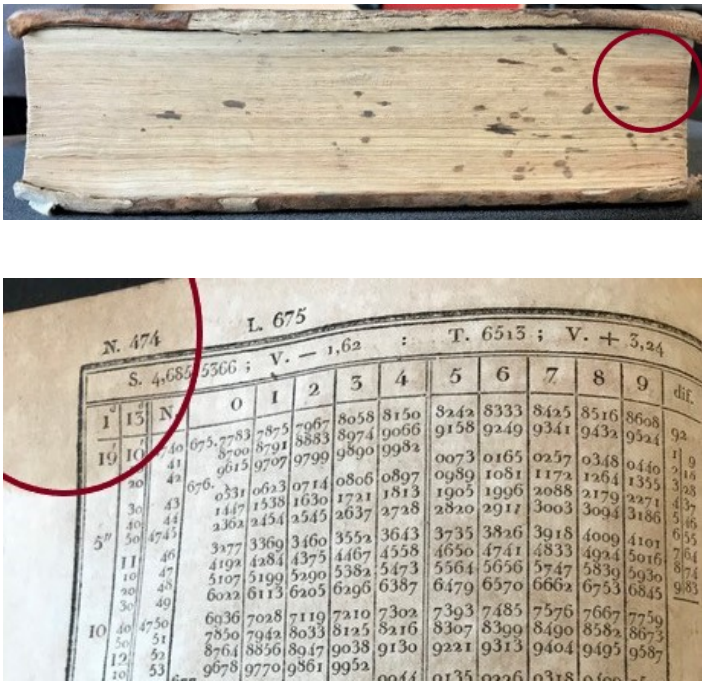


Figure 1 : Table logarithmique recueillie par M. Ducharme présentant des marques d'usure plus importantes sur les premières pages.

Cette découverte mise de côté pendant plusieurs années, ce n'est qu'en 1938 que l'ingénieur et physicien américain Frank Benford arrivera au même résultat après avoir répertorié des dizaines de milliers de données. Celui-ci pensera être le premier à l'initiative de cette loi, et c'est pour cette raison que la *loi de Newcomb-Benford* se fera plus généralement appelée *loi de Benford*.

Cette loi nous dit que, dans une liste de données arbitraires, la probabilité d'avoir le chiffre i comme premier chiffre significatif est de $\log_{10}(1 + \frac{1}{i})$.

PCS	1	2	3	4	5	6	7	8	9
Benford	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Tableau 1 : Répartition du premier chiffre significatif selon la loi de Newcomb-Benford.

2. Défini en 1985 par Marie-Paule Lecoutre (*source*).

Nous retrouvons cette loi dans énormément de domaines comme les mathématiques, l'environnement, la finance, la physique, etc, plus précisément sur des données telles que la longueur des fleuves, la population des villes dans un pays, des déclarations de revenus, etc.

Notons cependant qu'il existe des cas où les données ne suivent pas cette loi, notamment des données dites non naturelles qui seraient influencé par la pensée humaine (nombres premiers, nombres générés par des humains, etc).

Expérimentation sur différents jeux de données.

Après avoir pris connaissance de la **loi de Newcomb-Benford**, il serait intéressant de la mettre en pratique sur différents jeux de données.

La suite de Fibonacci.

Intéressons-nous dans un premier temps à la suite de Fibonacci.

Cette suite est une suite d'entiers dans laquelle chaque terme est la somme des deux termes qui le précèdent. Sa formulation est la suivante :

$$F_0 = 0, F_1 = 1, \text{ et } \forall n \geq 2, F_n = F_{n-1} + F_{n-2}.$$

Nous commençons par recueillir les 1000 premiers termes de la suite de Fibonacci, pour extraire le premier chiffre significatif de chacun de ces nombres.

Par la suite nous calculons la répartition de chaque chiffre significatif et obtenons l'histogramme suivant :

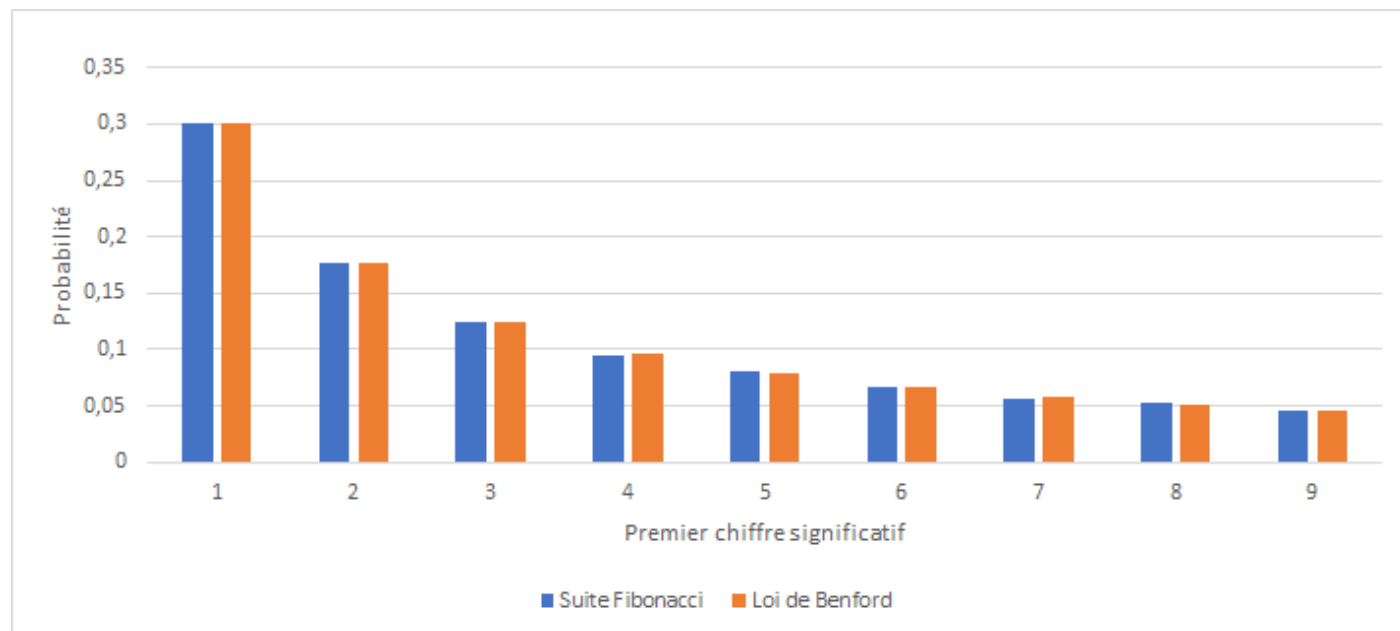


Figure 2 : Histogramme de la répartition du 1er chiffre significatif de la suite de Fibonacci en comparaison avec la loi de Benford.

Il semblerait que la répartition des chiffres significatifs des 300 premiers nombres de la suite de Fibonacci suive la **loi de Newcomb-Benford**.

Nombres extraits d'un magazine et d'un journal.

Dans un second temps, nous relevons les prix présents dans un magazine de mobilier de la marque *AMPM*, ainsi que tous les nombres répertoriés dans un journal *Les ECHOS*. Nous récoltons environ 300 nombres par magazine et, de la même façon qu'énoncé précédemment, calculons la répartition des chiffres significatifs de ces nombres.

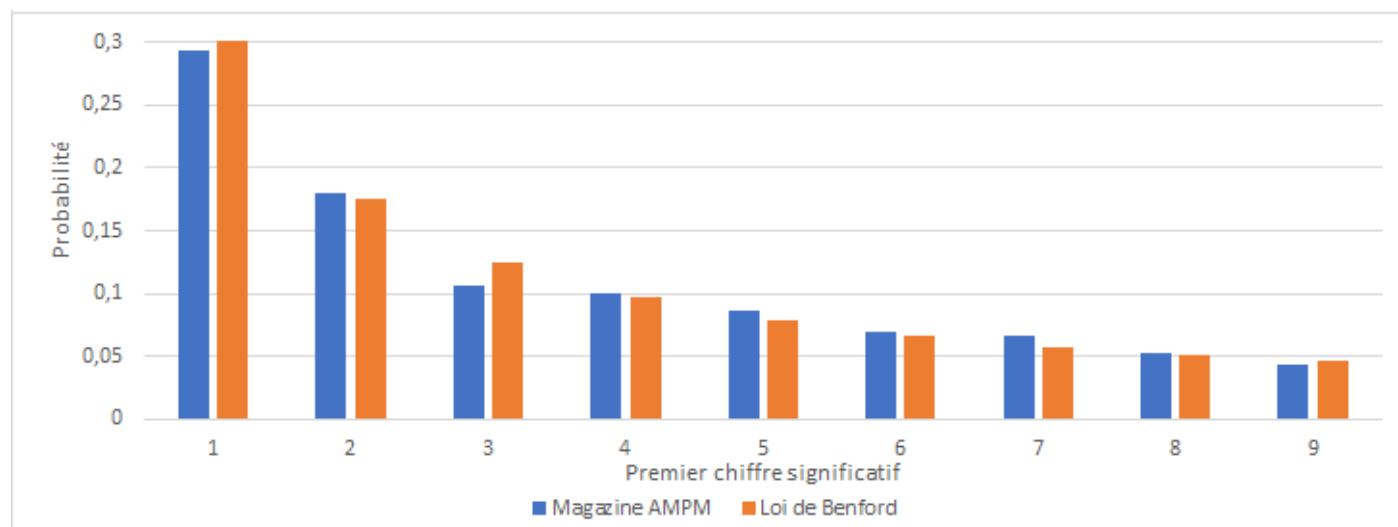


Figure 3 : Histogramme de la répartition du 1er chiffre significatif des prix du magazine AMPM en comparaison avec la loi de Benford.

La répartition des chiffres significatifs des prix du magazine *AMPM* paraît fortement similaire à celle de la **loi de Newcomb-Benford**. Nous constatons tout de même une légère différence pour le chiffre 3.

Observons maintenant la répartition des données issues du journal *Les ECHOS*.

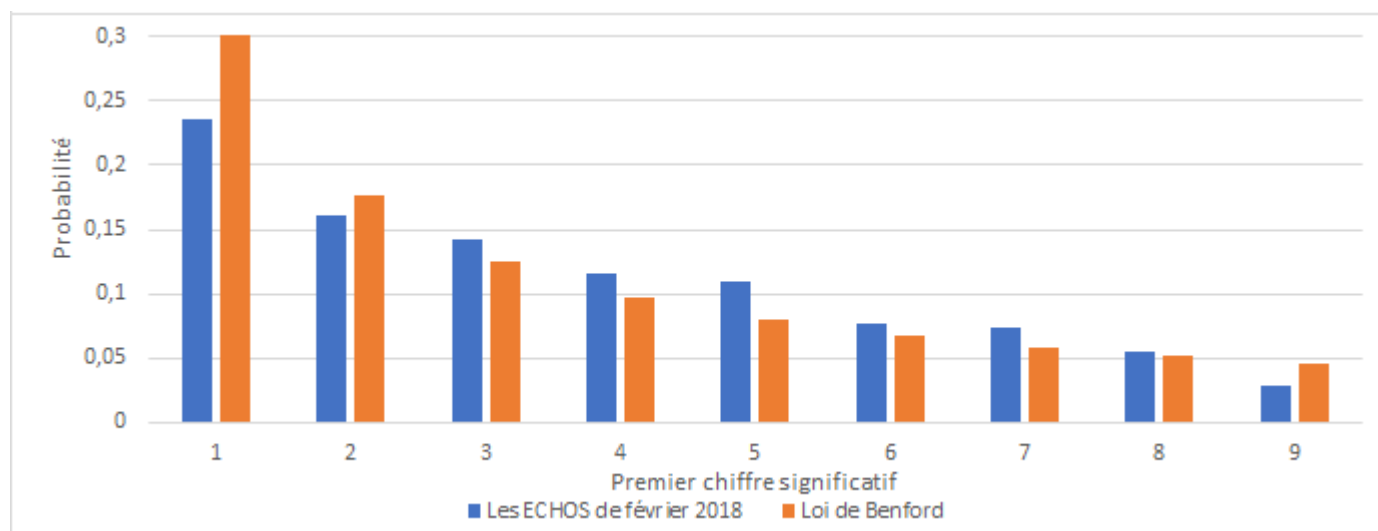


Figure 4 : Histogramme de la répartition du 1er chiffre significatif des nombres du journal LES ECHOS en comparaison avec la loi de Benford.

Nous remarquons ici la même tendance décroissante. Cependant les proportions des chiffres significatifs entre les données du journal et celles de la **loi de Newcomb-Benford** sont relativement différentes.

Population des villes de France.

Dans ce paragraphe nous nous intéressons à la population des villes françaises. À l'aide des données de l'*INSEE*, nous répertorions environ 35000 premiers chiffres significatifs et regardons leur répartition.

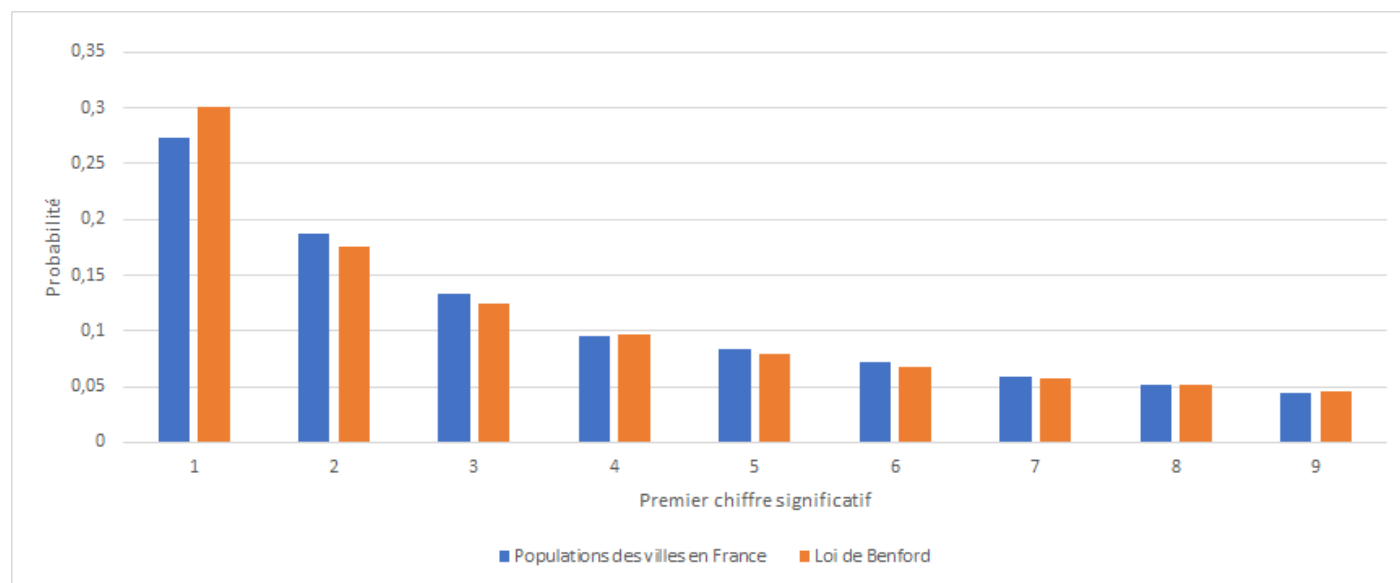


Figure 5 : Histogramme de la répartition du 1er chiffre significatif de la population française en comparaison avec la loi de Benford.

Ici les répartitions sont fortement ressemblantes, c'est aussi le cas pour de nombreuses données démographiques naturelles. Nous aurions pu également analyser les codes postaux, la longueur des rivières ou encore la distance des villes de France à Paris.

Passage journalier de vélos dans l'allée Beracasa à Montpellier.

La ville de Montpellier étant en pleine transition écologique, elle ouvre de plus en plus l'accès aux vélos sur ses routes. Pour en mesurer l'impact, elle a mis en place des éco-compteurs dans plusieurs rues. Les données issues de ces compteurs sont en libre accès, nous nous sommes donc intéressés au nombre de passages journaliers de vélos dans l'allée Beracasa sur une année.

Nous obtenons la répartition suivante :

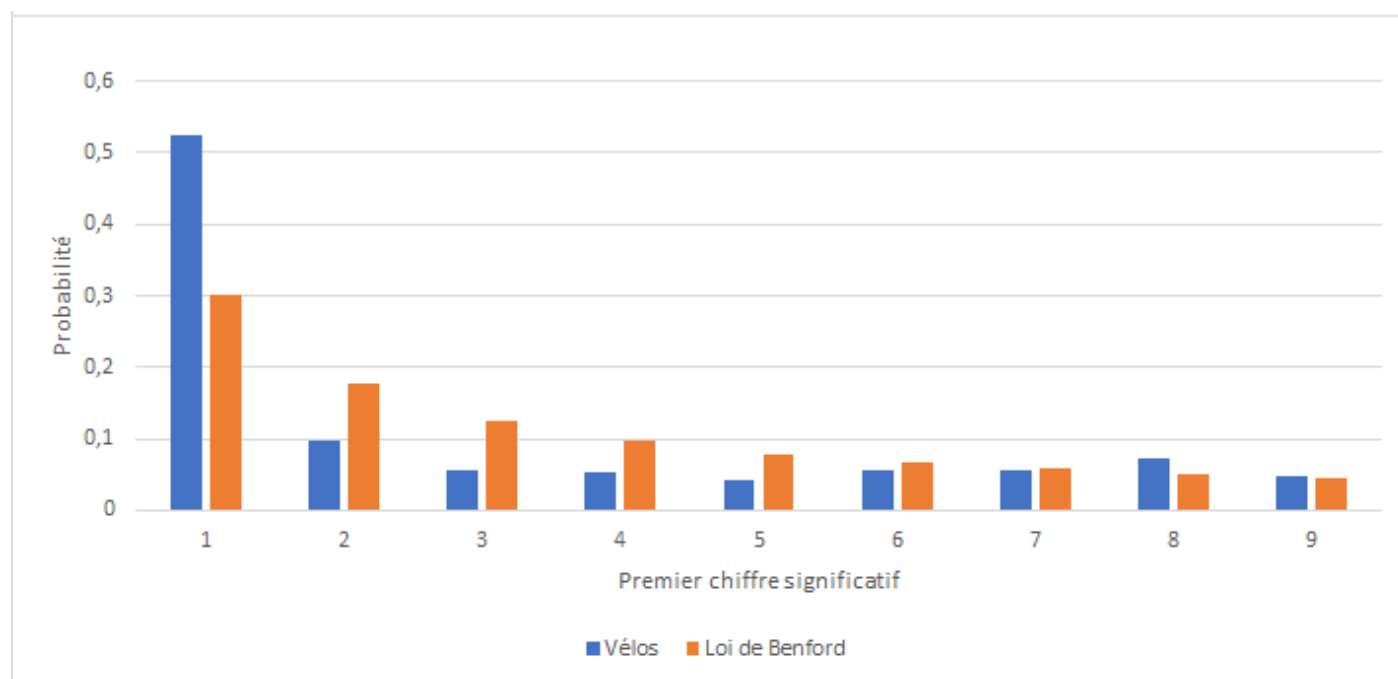


Figure 6 : Histogramme de la répartition du 1er chiffre significatif du passage journalier de vélos en comparaison avec la loi de Benford

Dans ce cas la proportion du chiffre 1 est de plus de 50% contre 30% pour la **loi de Newcomb-Benford**. La différence de répartition des chiffres 2, 3, 4, 5 est aussi notable, elle est même environ 2 fois moins élevée.

Visuellement, nous pourrions penser que la répartition de ces données ne suit pas la **loi de Newcomb-Benford**. Il est courant de ne pas retrouver la loi de Newcomb-Benford dans des données brutes comme celles-ci, on la retrouve empiriquement plus souvent dans des données dites de **deuxième génération** comme des sommes ou des produits. Ceci a été démontré par Jeff Boyle en 1994.

Nombres générés par les humains

Dans ce paragraphe nous tentons de reproduire à moindre échelle l'expérience de Theodore Preston Hill en 1988. Dans le cadre de son expérience le professeur Ted HILL demande à ses élèves (742) d'écrire un nombre de 6 chiffres au hasard sur un bout de papier, il recense ensuite le premier chiffre significatif de chacun de ces nombre dans le but de les comparer à la loi de Benford et à la répartition uniforme.

Notre expérience partageant le même objectif que celle de HILL, est basée sur un protocole légèrement différent. N'ayant pas une troupe d'élèves à disposition nous avons recueilli un total de 300 nombres. Ces

300 nombres ont été obtenus de plusieurs manières, via des sondages sur internet ou réseaux sociaux, en demandant directement à des personnes rencontrées au hasard, notre famille ou nos amis.

D'après le biais d'équiprobabilité cité plus haut, si les nombres recensés pendant les expériences ont réellement été données de façon aléatoire les répartitions des premiers chiffres significatifs devraient être comparable à une loi uniforme.

Comparons les répartitions obtenues durant les deux expériences :

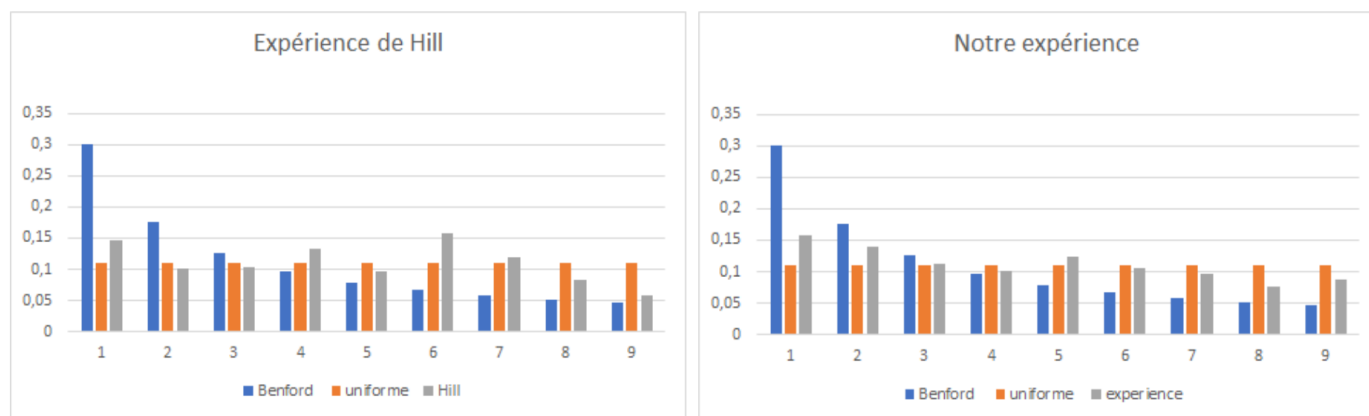


Figure 7 : Comparaison des histogrammes de la répartition du 1er chiffre significatif de l'expérience de Hill avec la loi de Benford et la loi uniforme puis de notre expérience avec ces deux mêmes lois.

À première vue, dans les deux expériences la répartition du premier chiffre significatif ne semble pas suivre la loi de Newcomb-Benford (le chiffre 1 n'apparaît clairement pas aussi souvent par exemple), elles semblent cependant plus proche de la loi uniforme sans tout de même y correspondre parfaitement. Plusieurs facteurs pourraient expliquer ces différences entre l'uniforme et les expériences, celui qui revient souvent est qu'un nombre donné au hasard par un humain est souvent influencé par son expérience, même inconsciemment. Par exemple sa date d'anniversaire, un évènement marquant ou son chiffre préféré. Le fait d'avoir recueilli nos nombres par internet à aussi pu influencer le choix des personnes concernées. Un facteur psychologique est donc à prendre en considération pour approfondir la conclusion.

Après avoir observé ces quelques jeux de données, nous étions en mesure de dire si ces données semblaient ou non suivre la loi de Newcomb-Benford, le problème qui en découle est qu'une simple observation n'est pas très fiable, difficile de prendre une décision sur un constat visuel. En effet, se tromper dans l'interprétation peut entraîner deux types d'erreur, la première étant de faussement déceler une fraude (ce que nous appellerons **le risque de première espèce**) et la deuxième de laisser passer une fraude. Ces erreurs ont un coût pour l'institut qui essaye de les réprimer, celui d'engager des démarches de détectations approfondies inutiles ou de ne pas percevoir les taxes dues dans le cas de la fraude fiscale par exemple.

Le but est donc de minimiser le coût que peuvent engendrer les erreurs sus-mentionnées, pour ce faire l'utilisation d'outils scientifiques est de rigueur. Les outils que nous utiliserons dans la suite sont les tests d'adéquations, ces tests servent à vérifier si un ensemble de nombre suit ou non une loi de probabilité donnée (pour nous c'est la loi de Newcomb-Benford).

Tests

Généralités sur les tests

Un test d'hypothèse (ou test statistique) est un procédé d'inférence statistique ayant pour but de fournir une règle de décision permettant ainsi, à partir de l'étude d'un ou plusieurs échantillons de données, d'indiquer si une hypothèse statistique concernant une population doit être acceptée ou rejetée.

Nous distinguons deux classes de tests :

- Les tests paramétriques sont l'étude de la moyenne, variance, ou de la fréquence des observations issues d'une distribution a priori paramétrée. Ils nécessitent un modèle à fortes contraintes (normalité des distributions ou approximation normale pour de grands échantillons). Ces hypothèses sont d'autant plus difficiles à vérifier que les effectifs étudiés sont plus réduits.
- Les tests non paramétriques sont l'étude des rangs des observations issues d'une distribution non paramétrée mais quelconque. Ce sont des tests dont le modèle ne précise pas les conditions que doivent remplir les paramètres de la population dont a été extrait l'échantillon. Il n'y a donc pas d'hypothèse de normalité au préalable.

Lorsque les conditions nécessaires sont valides, les tests paramétriques sont plus puissants que les tests non paramétriques. Les tests non paramétriques s'utilisent dès lors que les conditions d'application des autres méthodes ne sont pas satisfaites, même après d'éventuelles transformations de variables, et peuvent être employés pour des échantillons de taille très faible.

Comme nous l'avons précédemment énoncé, une inspection visuelle à elle seule ne permet pas d'affirmer ou infirmer si un jeu de données suit la loi de Newcomb-Benford. L'outil statistique permettant de le vérifier est le test d'adéquation à la loi de Newcomb-Benford.

Les tests d'adéquation servent à tester si un échantillon est distribué selon une loi de probabilité préalablement choisie. Ils permettent de décider, avec un certain seuil d'erreur, si les écarts présentés par l'échantillon par rapport aux valeurs théoriques sont dus au hasard, ou si au contraire ils sont significatifs.

Hypothèse nulle et hypothèse alternative

Un test statistique étudie deux hypothèses opposées concernant une population : l'hypothèse nulle et l'hypothèse alternative.

L'hypothèse nulle, notée H_0 , est l'hypothèse que l'on souhaite contrôler, elle repose sur le fait de dire qu'il n'existe pas de différence entre les paramètres comparés ou que la différence observée n'est pas significative et résulte des fluctuations d'échantillonnage.

À partir des échantillons de données, un test statistique permet de déterminer si on peut rejeter l'hypothèse nulle. La p -valeur sert de détermination. Si la p -valeur est inférieure au seuil de signification (appelé α), l'hypothèse nulle peut être rejetée.

L'hypothèse alternative, notée H_1 , affirme qu'un paramètre de la population est plus petit, plus grand ou différent de la valeur hypothétisée dans l'hypothèse nulle. Elle peut être vue comme la négation de H_0 et est équivalente à dire que H_0 est fausse. La décision de rejeter H_0 signifie que H_1 est réalisée ou que H_1 est vraie.

On pense souvent à tort que les tests d'hypothèses statistiques visent à choisir l'hypothèse la plus probable parmi H_0 et H_1 . Néanmoins l'hypothèse nulle est formulée dans le but d'être rejetée. Le seuil de signification fixé est bas (généralement 0.05), et lorsque l'hypothèse nulle est rejetée cela prouve statistiquement que l'hypothèse alternative est vraie. En revanche, si l'hypothèse nulle ne peut être rejetée aucune preuve statistique ne montre que celle-ci est vraie. La raison est qu'il n'y a pas de valeur fixée assurant que la probabilité d'accepter à tort l'hypothèse nulle est petite.

Finalement, la décision d'accepter l'hypothèse nulle n'est pas équivalente à dire que H_0 est vraie et que H_1 est fausse, mais cela traduit uniquement l'idée selon laquelle il n'y a pas d'évidence nette pour que H_0 soit fausse. Un test conclu donc à rejeter ou à ne pas rejeter l'hypothèse nulle mais jamais à l'accepter directement.

Dans la suite de notre projet, nous nous intéresserons aux données fiscales de 20 régions italiennes entre 2007 et 2011. Nous réaliserons donc pour chacune de ces régions et chacune des années le test d'hypothèse suivant : H_0 : "La répartition du premier chiffre significatif suit la loi de Newcomb-Benford" contre H_1 : "La répartition du premier chiffre significatif ne suit pas la loi de Newcomb-Benford".

Les risques d'erreurs

Notons alors qu'aucun test d'hypothèse n'est fiable à 100%, un test étant basé sur des probabilités, il existe toujours un risque de tirer une mauvaise conclusion. Lorsqu'un test d'hypothèse est effectué, nous pouvons observer deux types d'erreurs, l'erreur de Type I dite erreur de première espèce et l'erreur de Type II dite erreur de seconde espèce. Les risques de ces deux erreurs sont inversement proportionnels et sont déterminés par le seuil de signification (ou région critique) et la puissance du test. Il est important de déterminer l'erreur qui présente les conséquences les plus graves dans notre cas avant de définir le risque que nous accepterons pour chaque erreur.

L'erreur de Type I consiste à rejeter l'hypothèse nulle alors que celle-ci est vraie. La probabilité de commettre une erreur de première espèce est représentée par α , qui désigne le seuil de signification défini pour le test d'hypothèse. Un niveau α de 0.05 indique que nous sommes disposé à avoir 5% de chances de rejeter l'hypothèse nulle à tort. Pour réduire ce risque, il est possible d'utiliser une valeur α plus faible. Cependant, cela implique d'être moins à même de détecter une vraie différence si celle-ci existe vraiment.

Dans notre contexte, l'erreur de Type I consiste à affirmer que les données ne suivent pas la loi de Newcomb-Benford alors que c'est le cas, soit faussement identifier une fraude.

L'erreur de Type II repose sur le fait de ne pas rejeter l'hypothèse nulle alors que celle-ci est fausse. La probabilité de commettre une erreur de seconde espèce est notée β , et dépend de la puissance du test. Il est possible de réduire le risque de deuxième espèce en faisant en sorte que le test soit suffisamment puissant. Pour cela, il est nécessaire que l'effectif d'échantillon soit suffisamment grand pour permettre la détection d'une différence réelle.

La probabilité de rejeter l'hypothèse nulle à tort vaut $1 - \beta$, il s'agit de la puissance du test.

Décision d'après l'échantillon	Réalité sur la population	
	H_0 est vraie	H_0 est fausse
Ne pas rejeter H_0	Décision juste (probabilité = $1 - \alpha$)	Erreur de 2 ^{ème} espèce : acceptation de H_0 alors que celle-ci est fausse (probabilité = β)
Rejeter H_0	Erreur de 1 ^{ère} espèce : rejet de H_0 alors que celle-ci est vraie (probabilité = α)	Décision juste (probabilité = $1 - \beta$)

Tableau 2 : Risques d'erreurs

Au sujet de la loi de Newcomb-Benford, notons que le test d'adéquation le plus populaire est le test du khi-deux de Pearson dont la puissance, associée au risque d'erreur de Type II est relativement faible. C'est pourquoi, d'autres tests ont été mis en place récemment. L'ensemble de ces tests seront développés par la suite. G. Ducharme, S. Kaci et C. Vovor-Dassu ont introduits de nouveaux tests d'adéquation pour cette loi, basés sur le principe des tests lisses. Ils ont également comparés ces tests aux meilleurs tests existants et ont montrés qu'ils seraient globalement plus performants.

Test du Khi-Deux

Les tests du χ^2 sont des tests d'hypothèses statistiques non-paramétriques. Ceux-ci permettent de comparer la distribution observée dans un échantillon statistique avec une distribution théorique (*test d'ajustement*), à tester si deux caractères d'une population sont indépendants (*test d'indépendance*) et à tester si des échantillons sont issus d'une même population (*test d'homogénéité*). Le test lit l'écart critique dans la table de la loi du chi-deux.

Le déroulement du test se procède en 5 étapes :

- 1) On calcule les effectifs théoriques (n_{pj})
- 2) On calcule la valeur observée de la variable de test :

$$\chi^2 = \sum_{j=1}^k \frac{(n_j - n_{pj})^2}{n_{pj}}$$

- 3) On cherche la valeur critique χ_a^2 dans la table de la loi du χ^2 à $k - 1$ degrés de liberté.
- 4) Si $\chi_a^2 < \chi^2$, on accepte l'hypothèse H_0 ("la distribution observée est conforme à la distribution théorique" avec un risque d'erreur α), sinon on la rejette.
- 5) Il faut vérifier que $n_{pj} \geq 5$ pour tout j .

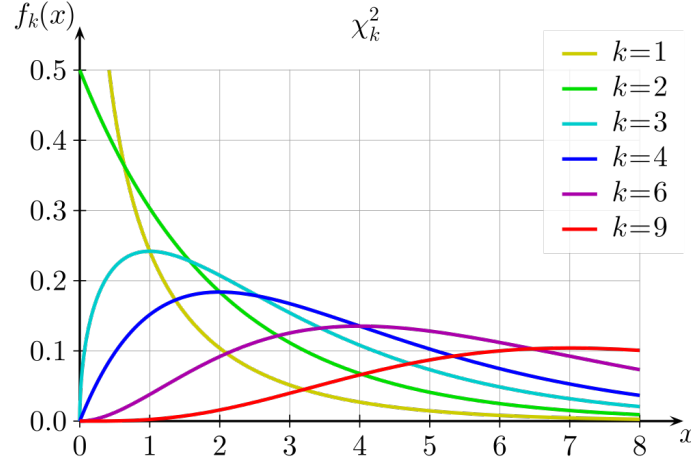


Tableau 3 : Densité de la loi du Chi-Deux en fonction du nombre de degrés de liberté.

Rappel de la loi du χ^2 à n degrés de liberté si elle est absolument continue, et admet pour densité :

$$f(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{\frac{n}{2}-1} & \text{si } x > 0, \\ 0 & \text{sinon,} \end{cases}$$

où X admet alors pour espérance et variance $E(X) = n$ et $V(X) = 2n$.

Test de Freedman-Watson

nbsp; Le test de Freedman est utilisé pour des distributions discrètes, il permet de comparer la distribution du premier chiffre significatif d'un échantillon de données avec la distribution de Benford et ainsi affirmer si les données sont conformes à la loi de Newcomb-Benford.

Spécifiquement, la statistique de test est donnée par :

$$U^2 = \frac{n}{9 \cdot 10^{k-1}} \cdot \left[\sum_{i=10^{k-1}}^{10^k-2} \left(\sum_{j=1}^i (f_j^o - f_j^e) \right)^2 - \frac{1}{9 \cdot 10^{k-1}} \cdot \left(\sum_{i=10^{k-1}}^{10^k-2} \sum_{j=1}^i (f_i^o - f_i^e) \right)^2 \right],$$

avec f_i^o la fréquence observée du chiffre i et f_i^e la fréquence attendue du chiffre i .

Notons que de plus grands écarts entre les fréquences conduisent à un plus grand U^2 , ce qui rend le rejet plus probable. Ce test est reconnu comme plus performant que d'autres tests, et a même été recommandé par la statisticienne M. Lesperance ainsi que ses collaborateurs (2016), puis également par Joenssen (2014).

Tests lisses pour la Loi de Newcomb-Benford

nbsp; La famille des tests lisses introduite par Neyman¹ s'applique à des données autant discrètes que continues. Il est spécifique à la loi de probabilité en H_0 .

Il existe deux théorèmes essentiels qui permettent de construire une famille de tests pour l'hypothèse nulle $H_0 : X \sim f(\cdot)$. Ici $f(\cdot)$ est la densité de la loi de Newcomb-Benford.

Le premier nous théorème nous dit :

Theorem 2..1. Soit X_1, \dots, X_n des copies indépendantes d'une variable aléatoire X de densité $f(\cdot)$ par rapport à une mesure dominante ν . Soit $\{h_0(\cdot) \equiv 1, h_k(\cdot), k = 1, 2, \dots\}$ une suite de fonctions orthonormales par rapport à $f(\cdot)$; plus précisément, $\int h_k(x)h_{k'}(x)f(x)d\nu(x) = \delta_{kk'}$, la fonction delta de Kronecker. Soit $U_k = n^{-1/2} \sum_{i=1}^n h_k(X_i)$ et pour un entier $K \geq 1$, soit $T_K = \sum_{k=1}^K U_k^2$. Alors sous H_0 , $T_K \xrightarrow{L} \chi_K^2$, la loi khi-deux à K degrés de liberté, et un test de niveau asymptotique α rejette H_0 si la valeur observée de T_K dépasse $x_{K,1-\alpha}^2$, le quantile d'ordre $1 - \alpha$ de cette loi χ_K^2 .

Nous avons donc nos statistiques de test, les T_K . Il faut maintenant exprimer les h_k . Dans la suite l'indice 0 dénote un opérateur probabiliste calculé sous $H_0 : X$ suit $f(\cdot)$.

D'après Ducharmes & Collab. et le théorème suivant :

Theorem 2..2. Soit $\mu_k = \mathbb{E}_0(X^k), k \geq 0$. Soit aussi la matrice $\mathbf{M}_k = [\mu_{i+i'}]_{i,i'=0,\dots,k-1}$, le vecteur $\boldsymbol{\mu}_k = (\mu_k, \mu_{k+1}, \dots, \mu_{2k-1})^T$ et la constante $c_k = \mu_{2k} - \boldsymbol{\mu}_k^T \mathbf{M}_k^{-1} \boldsymbol{\mu}_k$. Alors les polynômes

$$h_k(x) = c_k^{-1/2} (x^k - (1, x, x^2, \dots, x^{k-1}) \mathbf{M}_k^{-1} \boldsymbol{\mu}_k)$$

satisfont la condition du Théorème 2..1.

Nous avons pour $0 < k \leq 5$, les h_k suivants :

$$\begin{aligned} h_1(x) &= -1.3979 + 0.4063x \\ h_2(x) &= 2.2836 - 1.6128x + 0.18247x^2 \\ h_3(x) &= 4.0815 + 4.5719x - 1.2053x^2 + 0.0862x^3 \\ h_4(x) &= 8.0795 - 12.0946x + 5.1951x^2 - 0.8249x^3 + 0.0431x^4 \\ h_5(x) &= -18.1064 + 33.1385x - 19.7207x^2 + 5.0168x^3 - 0.5665x^4 + 0.0233x^5 \end{aligned}$$

Nous définissons

$$\hat{K} = \arg \max_{1 \leq k \leq K_{\max}} \{T_k - k \log(n)\}$$

et la statistique de test $T_{\hat{K}}$ qui $\xrightarrow{L} \chi_1^2$ sous H_0 .

Dans la suite nous appliquerons les test $T_{\hat{K}}$ et T_2 qui d'après Ducharme & Collab sont les plus performants.

Tests sur la région d'Abruzzo

ï..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.292	0.187	0.105	0.085	0.089	0.056	0.075	0.052	0.059	0.716	0.135	0.396	0.332
2008	0.295	0.170	0.111	0.089	0.098	0.059	0.062	0.056	0.059	0.870	0.220	0.465	0.309
2009	0.289	0.164	0.111	0.102	0.079	0.062	0.066	0.056	0.072	0.637	0.146	0.103	0.091
2010	0.298	0.154	0.121	0.092	0.085	0.069	0.066	0.052	0.062	0.918	0.115	0.331	0.211
2011	0.318	0.144	0.115	0.095	0.079	0.062	0.079	0.059	0.049	0.746	0.248	0.443	0.380

Tests sur la région de Basilicata

ï..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.290	0.160	0.130	0.061	0.122	0.076	0.061	0.023	0.076	0.267	0.289	0.708	0.428
2008	0.282	0.168	0.107	0.107	0.084	0.092	0.031	0.061	0.069	0.718	0.633	0.558	0.314
2009	0.290	0.160	0.122	0.099	0.061	0.099	0.053	0.053	0.061	0.894	0.225	0.647	0.388
2010	0.290	0.168	0.115	0.092	0.076	0.122	0.023	0.053	0.061	0.287	0.745	0.785	0.493
2011	0.305	0.160	0.099	0.115	0.053	0.122	0.046	0.046	0.053	0.348	0.505	0.874	0.611

Application des tests

Dans ce paragraphe nous traitons des données sur la fiscalité italienne de 20 régions, ces données nous serviront à appliquer différents tests (classique ou lisse). Chacun des tests teste l'hypothèse nulle "la répartition du premier chiffre significatif suit la loi de Newcomb-Benford" contre l'hypothèse alternative "la répartition du premier chiffre significatif ne suit pas la loi de Newcomb-Benford". L'objectif ici est de chercher à savoir si certaines régions d'Italie modifient ou non leurs chiffres, mais aussi de comparer les résultats de nos différents tests.

Ici nous retenons le test du khi-deux qui est le plus connu et utilisé, il semblerait cependant qu'il fasse parti des moins performants (Ducharmes & Collab. 2020), le test de Freedman-Watson explicité plus haut. Pour ce qui est des tests lisses nous appliquerons les test $T_{\hat{K}}$ et T_2 comme sus-mentionné.

Les données se présentent comme ci-après ## Application a des données fiscales italiennes

Tests sur la région de Calabria

ï..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.298	0.152	0.132	0.098	0.073	0.073	0.068	0.046	0.059	0.816	0.328	0.481	0.270
2008	0.301	0.164	0.127	0.093	0.073	0.068	0.073	0.042	0.059	0.808	0.258	0.599	0.446
2009	0.298	0.164	0.120	0.100	0.064	0.076	0.071	0.049	0.059	0.764	0.481	0.378	0.270
2010	0.308	0.166	0.115	0.103	0.061	0.081	0.073	0.051	0.042	0.712	0.606	0.921	0.719
2011	0.301	0.166	0.117	0.095	0.073	0.086	0.064	0.054	0.044	0.933	0.379	0.814	0.521

Tests sur la région de Campania

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.309	0.160	0.096	0.094	0.067	0.102	0.051	0.054	0.067	0.006	0.201	0.026	0.063
2008	0.314	0.160	0.093	0.100	0.065	0.100	0.044	0.062	0.064	0.003	0.404	0.034	0.114
2009	0.310	0.176	0.091	0.091	0.071	0.093	0.051	0.060	0.058	0.068	0.090	0.096	0.236
2010	0.310	0.172	0.091	0.087	0.082	0.087	0.049	0.065	0.056	0.094	0.069	0.089	0.186
2011	0.310	0.172	0.087	0.083	0.087	0.078	0.060	0.058	0.064	0.101	0.007	0.024	0.106

Tests sur la région d'Emilia-romagna

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.296	0.201	0.103	0.078	0.069	0.069	0.057	0.066	0.060	0.482	0.010	0.132	0.398
2008	0.310	0.204	0.101	0.080	0.066	0.066	0.060	0.063	0.049	0.634	0.035	0.311	0.979
2009	0.305	0.201	0.103	0.072	0.075	0.060	0.060	0.060	0.063	0.430	0.005	0.090	0.574
2010	0.290	0.201	0.112	0.072	0.072	0.055	0.066	0.063	0.069	0.225	0.019	0.044	0.250
2011	0.299	0.195	0.115	0.072	0.060	0.069	0.063	0.072	0.055	0.384	0.019	0.128	0.423

Tests sur la région de Friuli-Venezia-Giulia

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.279	0.210	0.142	0.078	0.055	0.059	0.068	0.064	0.046	0.639	0.083	0.889	0.878
2008	0.289	0.220	0.138	0.078	0.046	0.073	0.055	0.064	0.037	0.439	0.347	0.840	0.602
2009	0.280	0.220	0.142	0.078	0.041	0.069	0.064	0.064	0.041	0.343	0.097	0.911	0.870
2010	0.303	0.206	0.142	0.073	0.050	0.073	0.060	0.060	0.032	0.590	0.091	0.731	0.441
2011	0.284	0.211	0.151	0.069	0.055	0.073	0.041	0.073	0.041	0.287	0.182	0.897	0.805

Tests sur la région de la Lazio

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.304	0.193	0.119	0.095	0.061	0.058	0.053	0.056	0.061	0.760	0.076	0.302	0.974
2008	0.310	0.188	0.127	0.093	0.069	0.050	0.066	0.042	0.056	0.823	0.177	0.566	0.654
2009	0.315	0.193	0.116	0.087	0.082	0.048	0.061	0.040	0.058	0.659	0.302	0.421	0.539
2010	0.320	0.190	0.119	0.093	0.077	0.053	0.056	0.034	0.058	0.690	0.338	0.370	0.332
2011	0.312	0.198	0.127	0.079	0.077	0.061	0.045	0.048	0.053	0.829	0.138	0.436	0.375

Tests sur la région de la Liguria

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.272	0.209	0.085	0.089	0.077	0.098	0.043	0.047	0.081	0.044	0.078	0.172	0.148
2008	0.268	0.213	0.089	0.081	0.055	0.111	0.055	0.034	0.094	0.001	0.597	0.044	0.073
2009	0.294	0.209	0.077	0.098	0.060	0.094	0.060	0.034	0.077	0.047	0.106	0.267	0.418
2010	0.289	0.196	0.094	0.102	0.047	0.094	0.072	0.030	0.077	0.043	0.570	0.285	0.320
2011	0.306	0.191	0.089	0.115	0.047	0.089	0.060	0.043	0.060	0.286	0.246	0.650	0.800

Tests sur la région de Lombardia

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.298	0.168	0.118	0.092	0.087	0.074	0.064	0.050	0.048	0.725	0.449	0.479	0.225
2008	0.297	0.175	0.115	0.089	0.085	0.073	0.070	0.045	0.051	0.287	0.163	0.403	0.197
2009	0.299	0.176	0.114	0.093	0.085	0.072	0.070	0.048	0.043	0.500	0.395	0.759	0.526
2010	0.296	0.172	0.119	0.096	0.082	0.071	0.069	0.054	0.043	0.792	0.456	0.490	0.258
2011	0.300	0.172	0.117	0.097	0.078	0.074	0.067	0.051	0.043	0.775	0.258	0.783	0.507

Tests sur la région de Marche

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.268	0.172	0.092	0.121	0.084	0.063	0.054	0.071	0.075	0.199	0.108	0.049	0.026
2008	0.285	0.172	0.092	0.113	0.084	0.071	0.059	0.054	0.071	0.593	0.163	0.208	0.122
2009	0.280	0.180	0.088	0.117	0.088	0.071	0.067	0.042	0.067	0.497	0.139	0.420	0.200
2010	0.285	0.180	0.096	0.096	0.109	0.054	0.071	0.038	0.071	0.269	0.473	0.429	0.239
2011	0.293	0.197	0.088	0.096	0.105	0.042	0.071	0.054	0.054	0.376	0.501	0.716	0.567

Tests sur la région de Molise

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.243	0.125	0.191	0.118	0.074	0.074	0.074	0.051	0.051	0.284	0.175	0.165	0.162
2008	0.221	0.110	0.213	0.118	0.059	0.103	0.044	0.074	0.059	0.007	0.923	0.047	0.033
2009	0.250	0.110	0.184	0.110	0.081	0.103	0.051	0.066	0.044	0.156	0.860	0.110	0.116
2010	0.228	0.132	0.169	0.118	0.096	0.081	0.066	0.037	0.074	0.233	0.543	0.079	0.051
2011	0.235	0.140	0.169	0.096	0.103	0.081	0.081	0.029	0.066	0.252	0.219	0.117	0.087

Tests sur la région de Piemonte

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.304	0.185	0.119	0.091	0.088	0.061	0.064	0.052	0.036	0.616	0.430	0.626	0.459
2008	0.300	0.186	0.112	0.096	0.085	0.065	0.065	0.044	0.047	0.739	0.418	0.999	0.963
2009	0.303	0.180	0.110	0.095	0.090	0.061	0.061	0.053	0.046	0.766	0.373	0.925	0.787
2010	0.304	0.185	0.111	0.094	0.086	0.055	0.055	0.054	0.041	0.826	0.261	0.626	0.360
2011	0.299	0.186	0.114	0.095	0.076	0.056	0.056	0.051	0.042	0.680	0.311	0.621	0.384

Tests sur la région de Puglia

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.306	0.194	0.116	0.089	0.089	0.043	0.054	0.047	0.062	0.751	0.305	0.589	0.848
2008	0.295	0.213	0.101	0.101	0.074	0.078	0.039	0.047	0.054	0.648	0.907	0.903	0.750
2009	0.291	0.217	0.101	0.097	0.074	0.078	0.035	0.062	0.047	0.509	0.623	0.939	0.828
2010	0.302	0.209	0.105	0.101	0.078	0.074	0.031	0.062	0.039	0.558	0.565	0.770	0.470
2011	0.310	0.202	0.112	0.101	0.070	0.074	0.035	0.054	0.043	0.827	0.739	0.670	0.382

Tests sur la région de la Sardegna

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.310	0.167	0.093	0.053	0.050	0.119	0.064	0.109	0.034	0.000	0.761	0.010	0.000
2008	0.310	0.172	0.095	0.106	0.069	0.093	0.077	0.024	0.053	0.048	0.659	0.956	0.772
2009	0.316	0.175	0.101	0.095	0.085	0.069	0.093	0.032	0.034	0.084	0.476	0.825	0.741
2010	0.318	0.175	0.095	0.090	0.093	0.061	0.095	0.034	0.037	0.042	0.129	0.962	0.816
2011	0.314	0.174	0.103	0.087	0.092	0.066	0.103	0.029	0.032	0.006	0.832	0.860	0.880

Tests sur la région de la Sicilia

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.287	0.195	0.123	0.077	0.085	0.079	0.051	0.049	0.054	0.798	0.235	0.904	0.706
2008	0.272	0.205	0.123	0.077	0.082	0.069	0.067	0.044	0.062	0.460	0.088	0.602	0.362
2009	0.279	0.203	0.123	0.077	0.079	0.072	0.064	0.051	0.051	0.817	0.035	0.830	0.574
2010	0.282	0.197	0.131	0.074	0.085	0.056	0.079	0.038	0.056	0.285	0.389	0.845	0.631
2011	0.290	0.195	0.126	0.079	0.077	0.067	0.069	0.038	0.059	0.674	0.172	0.793	0.737

Tests sur la région de Toscana

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.331	0.188	0.125	0.059	0.066	0.070	0.035	0.059	0.066	0.171	0.112	0.045	0.013
2008	0.328	0.195	0.105	0.080	0.070	0.059	0.042	0.059	0.063	0.526	0.169	0.086	0.713
2009	0.321	0.206	0.094	0.084	0.073	0.056	0.038	0.066	0.063	0.255	0.087	0.080	0.841
2010	0.341	0.195	0.101	0.084	0.066	0.059	0.042	0.045	0.066	0.346	0.184	0.043	0.358
2011	0.369	0.185	0.101	0.091	0.073	0.042	0.056	0.038	0.045	0.280	0.563	0.023	0.030

Tests sur la région de Trentino-alto

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.322	0.159	0.127	0.100	0.077	0.065	0.062	0.062	0.027	0.788	0.329	0.801	0.562
2008	0.300	0.168	0.126	0.099	0.081	0.048	0.075	0.060	0.042	0.836	0.451	0.966	0.822
2009	0.285	0.171	0.135	0.105	0.078	0.048	0.084	0.039	0.054	0.436	0.991	0.889	0.655
2010	0.300	0.168	0.120	0.114	0.078	0.057	0.072	0.048	0.042	0.935	0.534	0.922	0.968
2011	0.294	0.156	0.117	0.120	0.087	0.048	0.072	0.048	0.057	0.538	0.861	0.736	0.437

Tests sur la région de Umbria

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.38	0.207	0.130	0.065	0.109	0.054	0.011	0.022	0.022	0.210	0.786	0.030	0.021
2008	0.37	0.207	0.130	0.065	0.120	0.043	0.022	0.022	0.022	0.241	0.426	0.052	0.025
2009	0.37	0.207	0.130	0.076	0.120	0.043	0.022	0.022	0.011	0.196	0.701	0.022	0.017
2010	0.37	0.217	0.120	0.076	0.109	0.065	0.011	0.000	0.033	0.136	0.871	0.035	0.019
2011	0.38	0.217	0.109	0.087	0.109	0.054	0.011	0.011	0.022	0.154	0.683	0.018	0.013

Tests sur la région de valledaosta

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.270	0.230	0.095	0.081	0.122	0.054	0.081	0.041	0.027	0.708	0.633	0.728	0.902
2008	0.270	0.243	0.108	0.068	0.108	0.054	0.081	0.014	0.054	0.563	0.806	0.939	0.825
2009	0.284	0.203	0.122	0.054	0.149	0.054	0.068	0.027	0.041	0.486	0.606	0.799	0.870
2010	0.284	0.216	0.108	0.027	0.162	0.068	0.054	0.041	0.041	0.185	0.645	0.923	0.983
2011	0.284	0.176	0.135	0.041	0.135	0.081	0.054	0.027	0.068	0.501	0.984	0.908	0.654

Tests sur la région de veneto

i..Digit	X1	X2	X3	X4	X5	X6	X7	X8	X9	Chi2	Watson	T2	Tk_hat
2007	0.310	0.165	0.136	0.117	0.057	0.057	0.062	0.040	0.057	0.184	0.172	0.770	0.714
2008	0.313	0.167	0.139	0.107	0.062	0.062	0.065	0.041	0.043	0.619	0.268	0.666	0.374
2009	0.320	0.170	0.138	0.105	0.062	0.064	0.062	0.043	0.036	0.613	0.408	0.316	0.139
2010	0.325	0.172	0.138	0.103	0.067	0.062	0.055	0.053	0.024	0.297	0.527	0.094	0.042
2011	0.322	0.170	0.134	0.105	0.071	0.057	0.060	0.048	0.033	0.704	0.763	0.270	0.118

CAS COVID

Un chinois a affirmé qu'il n'y avait pas eu fraudes en Chine

Bibliographie

- V. GENEST, C. GENEST *La loi de Newcomb-Benford ou la loi du premier chiffre significatif* (2011)
- WIKIPEDIA L'ENCYCLOPÉDIE LIBRE *Loi de Benford* (dernière mise à jour 2021)
- T. P. HILL *Random-number guessing and the first digit phenomenon* (1988)
- G. R. DUCHARME, S. KACI, C. VOVOR-DASSU *Tests d'adéquations lisses pour la loi de Newcomb-Benford* (2017)
- Chi-deux :
- <http://www.bibmath.net/dico/index.php?action=affiche&quoi=.1/loichideux.html> https://fr.wikipedia.org/wiki/Test_du_%CF%87%C2%B2#Test_du_%CF%872_de_Pearson <http://www.bibmath.net/dico/index.php?action=affiche&quoi=.c/chideux.html> https://commons.wikimedia.org/wiki/File:Chi-square_pdf.svg?uselang=fr
- Freedman : D. JOENSSEN et T. MUELLERLEILE *Package R "BenfordTest"* (2015) <https://cran.r-project.org/web/packages/BenfordTests/BenfordTests.pdf> D. JOENSSEN (2014) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2545243 LESPERANCE M., REED W.J., STEPHENS M.A., TSAO C., WILTON B.,(2016) "*Assessing conformance with Benford's law : Goodness-of-fit tests and simultaneous confidence interval*"
- M. AUSLOOS, R. CERQUETI, T. A. MIR *Data science for assessing possible tax income manipulation : the case of Italy* (2017)
- Test lisse : BOULERICE B., DUCHARME G.R. (1997) "*Smooth test of goodness-of-fit for directional and axial data*", *Journal Multivariate Analysis* 60 : 154-175 G.R. DUCHARME, S. KACI, C. VOVOR-DASSU (2017) article *Tests d'adéquations lisse pour la loi de Newcomb-Benford* https://www.researchgate.net/publication/339641985_Smooths_Tests_of_Goodness-of-fit_for_the_Newcomb-Benford_distribution NEYMAN J. (1937) , "*Smooth tests for goodness-of-fit*", *Skand. Aktuarietidskr* 20 : 150-199.